AN EFFICIENT DOCKING ALGORITHM USING CONSERVED RESIDUE INFORMATION TO STUDY PROTEIN-PROTEIN INTERACTIONS

YUHUA DUAN[1*], BOOJALA V. B. REDDY[2] AND YIANNIS N. KAZNESSIS[1,2]
[1]*Department of Chemical Engineering and Materials Science, and*
[2]*Digital Technology Center,*
University of Minnesota, Minneapolis, MN 55455

ABSTRACT

Many protein-protein docking algorithms generate numerous possible complex structures with only a few of them resembling the native structure. The major challenge is choosing the near-native structures from the generated set. Recently it has been observed that the density of conserved residue positions is higher at the interface regions of interacting protein surfaces, except for antibody-antigen complexes, where a very low number of conserved positions is observed at the interface regions. In the present study we have used this observation to identify putative interacting regions on the surface of interacting partners. We studied 59 protein complexes, used previously as a benchmark dataset for docking investigations. We computed conservation indices of residue positions on the surfaces of interacting proteins using available homologous sequences and used this information to filter out from 55% to 88% of generated docked models, retaining near-native structures for further evaluation. We used a reverse filter of conservation score to filter out the majority of non-native antigen-antibody complex structures. For each docked model in the filtered subsets, we relaxed the conformation of the side chains by minimizing the energy with CHARMM. We then calculated the binding free energy using a generalized Born method and solvent accessible surface area calculations. Using the free energy along with conservation information and other descriptors used in the literature for ranking docking solutions, such as shape complementarity and pair-potentials, we developed a global ranking procedure that significantly improves the docking results by giving top ranks to the near-native complex structures.

1. Introduction

Predicting the structure of protein-protein complexes using computational methods has progressed substantially (Cherfils et al. 1993; Janin 1995; Shoichet et al. 1996; Sternberg et al. 1998; Camacho et al. 2002; Halperin et al. 2002; Smith et al. 2002). Numerous docking algorithms have been developed based on shape complementarity search algorithms (Katchalski-Katzir et al. 1992), such as PUZZLE (Helmer-Citterich et al. 1994), DOCK (Ewing et al. 2001), FTDock (Gabb et al. 1997), DOT (Mandell et al. 2001), and ZDOCK (Chen et al. 2003a). Since protein-protein docking is a hard problem to address due to the large number of degrees of freedom involved, some new techniques were introduced into docking procedures: HEX uses expansion of the molecular surface and electric field in spherical harmonics (Ritchie et al. 2000), BIGGER involves surface-implicit methods (Palma et al. 2000), AutoDock (Morris et al. 1998), DARWIN (Taylor et al. 2000; Gardiner et al. 2003), GAPDOCK (Gardiner et al. 2003) and GEMDOCK (Yang et al. 2004) employ genetic algorithms.

In principle, calculation of the free energy change upon binding of two proteins should allow determination of the native structure. Although the enthalpic part of the free energy can be calculated with some accuracy, the entropic contributions are not easy to calculate without resorting to semiempirical and less accurate calculations. Furthermore, the computational load can become too large, especially for unbound docking (starting with individual protein crystal structures) which can potentially involve large protein conformation changes. Heuristic criteria, such as shape complementarity and coarse-grained residue potentials have been used with relative success. Still, the main bottleneck is choosing the near-native structures from large sets of generated complexes based on a standard global ranking procedure that will bring the near-native structures at the top of the generated structures dataset.

Additional information has been used to better select near-native structures: HADDOCK (Dominguez et al. 2003) and TreeDock (Fahmy et al. 2002) use information based on chemical shift perturbation data resulting from NMR titration experiments or mutagenesis, whereas ConsDock (Paul et al. 2002) uses consensus analysis for protein-ligand interactions. ClusPro (Comeau, et al. 2004) has been implemented as a fully automated server.

Recent studies of protein complexes have tested the importance of factors, such as interface propensity of residues, accessible surface area, planarity, protrusion, packing energies and binding areas (Jones et al. 1996; Tsai et al. 1997; Larsen et al. 1998; Lo Conte et al. 1999). A test using averages of these factors as an indicator of protein binding sites showed about 66% success rate for 59 predictions (Jones et al. 1997).

There have also been several reports investigating the role of conservation of interfacial residues in naturally occurring protein complexes, using evolutionary tracing of conserved residues in homologous sequences and structures (Lichtarge et al. 2002; Glaser et al. 2003). Our recent analysis of well-resolved protein complexes indicated that the density of highly conserved residues is higher in protein-protein interface positions compared to the other positions of the protein surfaces (Reddy et al. 2004). We actually find that highly conserved positions

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **00 DEC 2004**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE
An Efficient Docking Algorithm Using Conserved Residue Information To Study Protein-Protein Interactions

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Chemical Engineering and Materials Science, and Digital Technology Center, University of Minnesota, Minneapolis, MN 55455

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2005 in Orlando, Florida. , The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**8**	

in surface regions of proteins involved in non-antibody-antigen complexes tend to be in interacting patches. On the other hand, for antibody-antigen complexes, a very low number of conserved positions is observed in the interface regions. This information can potentially assist in the selection of near-native structures. However, to our knowledge, no attempts have been made to use residue conservation information to filter and rank the docking solutions of protein complexes.

In this paper we describe our docking analysis and ranking of docked complex structures for 59 benchmark complexes (Chen et al. 2003b). We have used FTDock (Gabb et al. 1997; Moont et al. 1999) to generate 10,000 docked models for each of the complexes. We have then used conserved residue position information as a filter to reduce the number of docked structures. Besides filtering, we use conservation information to rank the remaining docked structures. We evaluate these approaches and report on the results.

In this paper we also report on our efforts to develop a global ranking scheme. For each docked model, we relax the conformation of the side chains by minimizing the energy with CHARMM and then calculate the binding free energy using a generalized Born method and the solvent accessible surface area. We finally develop a global ranking procedure so that the near-native structures rank at the top, using all available information from docking, free energy calculations and residue conservation information.

2. Methods

In Figure 1, we present a diagram with the steps that constitute our method. Briefly, for any two protein molecules A and B, we generate 10,000 structures using FTDock (Gabb et al. 1997; Moont et al. 1999). FTDock also calculates a shape complementarity value and a pair-potential value for these 10,000 model structures. We then calculate conservation indices for the surface positions of the proteins and also calculate the desolvation energy upon binding. Using these two properties, along with the shape complementarity and the pair potential we develop two filters to reduce the number of model structures to a number considerably lower than 10,000. We then use CHARMM to minimize the energy of the filtered structures and we calculate the free energy of binding. Finally, we use the ranks of the model structures for all the properties to generate a global ranking scheme, which improves our ability to pick near-native structures from the set of putative native structures. The methods are detailed as follows.

2.*1 Docking Calculations*

To generate model docked structures, we employed FTDock software package (Gabb et al. 1997; Moont et al. 1999); (http://www.bmm.icnet.uk/docking), which uses an efficient geometric recognition algorithm to identify molecular surface complementarity (Katchalski-Katzir et al. 1992). This method is based on a purely geometric approach and takes advantage of techniques applied in the field of pattern recognition. The geometric recognition algorithms include a digital representation of the proteins by 3D discrete functions for surface and the interior, a correlation function calculation using Fourier transformation that assesses the degree of molecular surface overlap and penetration upon relative shifts of the molecules in 3D, and a scan of the relative orientations of the molecules in 3D. From FTdock calculation, we obtained shape complimentarity rank and pair-potential rank.

2.2 Conservation of Residue Positions

In order to evaluate the extent of conservation of interacting positions on the surface of proteins we calculate conservation indices as follows:

2.2.1 Homologous sequences

The two protein sequences of each investigated complex were used to obtain their homologous sequences from SWALL, an annotated non-redundant protein sequence database (non-redundant swissprot + TrEMBL + TrEMBLnew), using FASTA3(www.ebi.ac.uk/fasta33/)

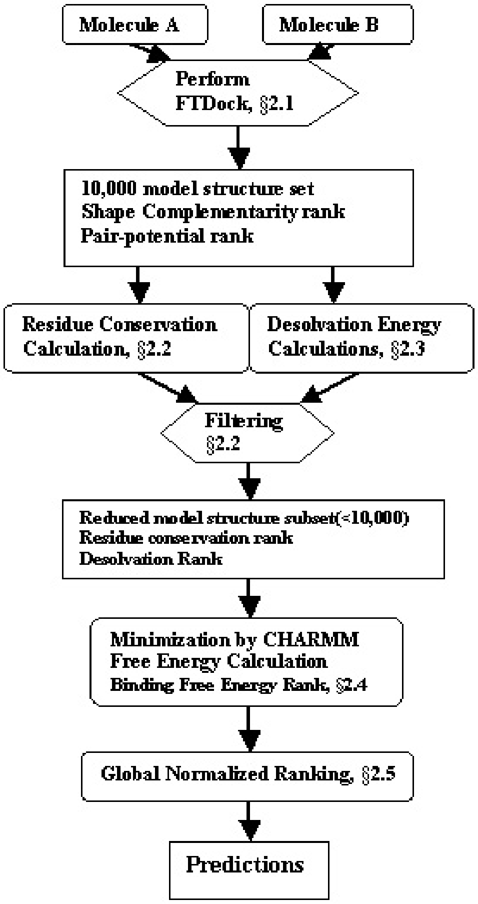
Molecule A
Molecule B
Perform FTDock, §2.1
10,000 model structure set
Shape Complementarity rank
Pair-potential rank
Residue Conservation Calculation, §2.2
Desolvation Energy Calculations, §2.3
Filtering §2.2
Reduced model structure subset(<10,000)
Residue conservation rank
Desolvation Rank
Minimization by CHARMM
Free Energy Calculation
Binding Free Energy Rank, §2.4
Global Normalized Ranking, §2.5
Predictions

Fig.1. Schematic representation of the procedure used sequence similarity search tool at the European Bioinformatics Institute. Homologous sequences with less than 30% gaps in the sequence and greater than 35% sequence identity to the parent sequence were used for analysis. If the evolutionary distance (described below) between any two sequences is less than 5% then we randomly removed one of the sequences from the homolog set. The remaining sequences were used for calculating the residue conservation index (described below).

2.2.2 Evolutionary Distance

Evolutionary distance among the sequences is calculated using the structure based amino acid substitution matrix (Gonnet et al. 1992). A similarity score S_{ii} for sequence *i* is calculated by summing the identical substitution values of the residues *a* and *b* from the substitution matrix *M(a,b)*. Similarly, score S_{jj} is calculated for sequence *j*. A similarity score S_{ij} between the sequences *i* and *j* is calculated using substitution matrix values of corresponding aligned residues between the two sequences. An evolutionary distance (ED_{ij}) between the two sequences is calculated using

$$ED_{ij} = \left[\left(\left(1-\frac{S_{ij}}{S_{ii}}\right)+\left(1-\frac{S_{ij}}{S_{jj}}\right)\right)/2\right]\times 100 \qquad (1)$$

2.2.3 Conservation Index of Residue Position

Evolutionary distances between the reference sequence and its homologues were used to calculate residue conservation index (CI_l) for each position *l* using the amino acid substitution matrix, similar to the amino acid variability or conservation used by Valdar and Thornton (Valdar et al. 2001). Conservation Index (CI_l) is a weighted sum of all pairwise similarities between all residues present at the position. The CI_l value is calculated using equation (2) in a given alignment and takes a value in the range [0,1].

$$CI_l = \frac{\sum_{i}^{N}\sum_{j>i}^{N} ED(s_i)\times ED(s_j)\times Mut(s_i(l), s_j(l))}{\sum_{i}^{N}\sum_{j>i}^{N} ED(s_i)\times ED(s_j)} \qquad (2)$$

where N is the number of homologous sequences in the alignment; $s_i(l)$ and $s_j(l)$ are the amino acids at the alignment position *l* of sequences s_i and s_j respectively; $ED(s_i)$ and $ED(s_j)$ are the average evolutionary distance of *s(i)* and *s(j)* from the remaining homologues. *Mut(a,b)* measures the similarity among the amino acids *a* and *b* as derived from amino acid substitution matrix *M(a,b)* and defined as:

$$Mut(a,b) = \frac{M(a,b) - M(a,b)low}{M(a,b)\max - M(a,b)low} \qquad (3)$$

where a, b are the pairs of amino acids at a given alignment position *l*. $M(a,b)_{low}$ is the lowest value in the substitution matrix (-5 in the Gonnet matrix (Gonnet et al. 1992)) and $M(a,b)_{max}$ is the maximum value among all the possible substitution pairs in that position. Thus the *Mut(a,b)* takes a value in the range [0,1].

Using PSA (Richmond et al. 1978; Sali et al. 1990), the solvent accessible surface area (SASA) of amino acids is calculated and used to identify surface residues and buried residues. We have then identified the top 8% and 17% percent of highly conserved residues, which have solvent accessibility greater than 25% of their total surface area.

For each complex, we add all conservation indices for each conserved position and use them to rank the complexes after filtering. In this case, two conservation ranks are obtained for group 1 and 2 respectively. We have observed (Reddy et al. 2004) that in the functionally interacting natural proteins, such as enzyme-inhibitor complexes, the number of conserved positions is significantly higher in the interface region than in the rest of the protein surface. We thus assigned high ranks to complexes that had a large number of conserved positions at the interacting interface. In the case of antigen-antibody complexes the interacting regions are highly variable, and we gave higher ranks to the models with low numbers of conserved positions.

2.3 Filters

2.3.1 Conservation position filter

Using homologous sequences we calculated conservation indices for each docked model using eq. 5. We have identified the top 8% (defined as group 1) and top 17% (defined as group 2) of highly conserved and well-exposed surface residues, in each polypeptide chain of the interacting complex.

We counted the total number of group 1 and group 2 positions in each modeled complex interface region. Using the group 1 and group 2 conservation positions as a filter, the total number of docked models are reduced. We selected only the models, which have at least 4 of group 1 positions or 6 of group 2 positions in the interface region of the enzyme-inhibitor model complexes. In the case of antigen-antibody complexes (for example 1JHL, 1KXQ, etc.) we have reversed the selection, limiting to 2 or less group 1 positions and 4 or less group 2 positions. We chose these cut-offs because we maximized the number of filtered docking solutions out of the 10,000 generated structures with the minimum number of near-native structures, as discussed in the Results section.

2.3.2 Filter II

A second filter was developed to lower the number of model structures further, using the average conservation rank along with other three ranks (shape-complementarity, pair-potential, and desolvation energy (described in the next section)). If the rank of a complex is worse than 1,200 in any of the four rankings then the corresponding model is filtered out of the set of putative near native structures. Filter II is performed with only three ranks if conservation information is not available as described in the results section.

2.4 Side-chain Relaxation and Binding Free Energy Calculation

Since the generated docked complexes have very strong side-chain overlap effects (atoms are very close to each other) we cannot calculate the binding energy correctly. Therefore, for each possible complex we perform energy minimization to reduce the side chain overlap effects. We employed CHARMM (Brooks et al. 1983) molecular mechanics simulation package for energy minimization. With CHARMM we built in the missed atoms and all hydrogen atoms, fixed all backbone atoms and let the side-chain atoms relax to the minimum internal energy. Minimization was stopped if the energy did not change by more than 0.1% of the total energy of the complex. We should note here that this step is particularly computationally intensive. We thus worked on only the filtered structures after using the calculated conservation indices.

Using the relaxed structures, we calculated the binding free energy. With some approximation, the free energy change can be divided into several terms (Dennis et al. 2002):

$$\begin{aligned}\Delta G &= \Delta G_{es} + \Delta G_{cav} + \Delta G_{bonding} + \ldots \\ &\approx \Delta G_{coulomb} + \Delta G_{pol} + \sum_{i} \sigma_i SASA_i + \Delta G_{bonding}\end{aligned} \tag{4}$$

These terms can be calculated separately: $\Delta G_{coulomb}$ and ΔG_{pol} can be calculated with the Generalized Born model with the Debye-Huckel approximation (Jayaram et al. 1998; Jayaram et al. 1999).

The desolvation energy term $\Sigma\sigma_k SASA_k$ can be calculated using the solvent accessible surface area for each residue ($SASA_k$). The weights (σ_k) for each residue are taken from the work of Wang and co-workers (Wang et al. 1995). For the binding interaction, we use van der Waals interaction of the form. The potential parameter A_{ij} and B_{ij} for each atom pair are taken from CHARMM force field (Brooks et al. 1983) and AutoDock (Morris et al. 1998). From the value of free energy ΔG, we calculated a new rank for all filtered possible complexes.

We also generated a rank based on only the desolvation term of the free energy, which is the only part of the free energy that can be calculated without relaxing the docked structures with minimization.

2.5 Global Normalized Ranking

Our goal is to determine an optimal ranking procedure for identifying near-native structures. We could use a weighted sum of all the calculated descriptors (shape-complementarity, pair-potential, CHARMM energy, binding free energy, desolvation energy, conservation indices) to produce a global rank for the filtered subset of docked models, but values of these properties are not in the same units and the weights are not universal and hard to optimize. In our algorithm, instead of using the real value of each descriptor, we used the rank of each property since they have the same meaning and can be summed together.

For each individual descriptor, a normalized ranking method is applied. The rank was obtained by finding the maximum (Vmax) and minimum (V_{min}) of their values and using the following equation.

$$NORM\_RANK_i = 1 + ANINT\left(\frac{V_i}{\frac{V_{max} - V_{min}}{N}}\right) \tag{5}$$

where V_i is the property value of complex i, and N is the total number of complexes after filtering. There may be some gaps if the difference between complexes is large, and several complexes can have the same rank number if their values are very close to one another. Nonetheless, this normalized method clearly reveals the difference among the complexes. Specifically for the binding free energy descriptor, we set the V_{max} equal to zero. If for a complex the binding free energy is greater than zero, we assign the highest rank (in our case is 10,000) to that complex.

The global score is simply obtained by average of all normalized ranks.

$$GLOBAL\_Score = \frac{1.0}{100 * M} \sum_{i}^{M} \sigma_i * NORM\_RANK_i \tag{6}$$

where M is the number of rank methods (descriptors), σ_i is the weights for descriptor i. Factor 100 is a scale factor which reduces the maximum of global_score to 100.

3. Results and Discussion

In order to test the usefulness of our filter and ranking methods, we applied our algorithms to a benchmark of 59 non-redundant protein complexes first used by Chen and co-workers (Chen et al. 2003b). This benchmark set includes 22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, 11 other complexes and 7 difficult test cases. This benchmark has been used by other groups to test their docking methods (Gray et al. 2003). Gottschalk and co-workers also used 21 complexes of this benchmark to test their scoring function of tightness of fit (Gottschalk et al. 2004). Since unbound-unbound docking (using single protein crystal structures as input) is more challenging than bound-bound docking (using the structures obtained from protein-complex crystals), we have carried out the unbound-unbound docking.

3.1 Analysis of FTDock performance

Using FTDock, we obtained 10,000 docked models and their ranks according to the correlation function of shape complementarity and pair potential (see section 2.1). For these 10,000 models, we calculated the root mean square deviation (RMSD) of Cα atoms of each model structure from the native structure. We then defined "hits" as the number of models having RMSD less than 4.5Å from the native structure. It can be seen from the results that there are 26 complexes with LRMSD less than 2.5Å, 15 complexes with LRMSD greater than 2.5Å but less than 3.5Å, and 8 complexes with LRMSD greater than 3.5Å. We are thus confident that FTDock can generate model complexes close to native structures. Nonetheless, for 5 complexes (1AVW, 1BQL, 1EFU, 1FIN, 1GOT), FTDock failed to generate near-native structures, as the LRMSDs for these complexes are greater than 4.5Å.

The rank based on shape complementarity predicts near native structures very poorly: the average rank of the LRMSD complexes is 4123, with only three of the 60 complexes registering ranks better than 100. It is thus clear that shape complementarity is not by itself an adequate means for choosing near-native structures.

The pair-potential rank did improve the ranks for 47 complexes out of the 60 cases. From the results, it can be observed that there are only 12 complexes with pair-potential ranking worse than shape complementarity. Nonetheless, ranks based on pair-potential do not have impressive predictive ability. For example, only 5 complexes (1BRC, 1BRS, 1PPE, 2MTA, 2SIC) have ranks less than 20 for the LRMSD model and another 3 complexes (1CGI, 1CHO, 2BTF) have ranks of LRMSD complexes less than 100. The rest have very high rank values.

3.2 Filters performance

First, we try to reduce the number of possible docked models from the generated 10,000, without filtering out the lower RMSD models. As described in section 2.3, we developed two filters based on residue conservation information. In the functionally interacting natural proteins, such as enzyme-inhibitor complexes, we gave higher ranks for the models with higher number of conserved positions in the interface region. In the case of antigen-antibody interactions the interacting regions are highly variable, and we gave higher ranks for the models with low numbers of conserved positions. With the conservation positions filter we reduced the number of complexes by about 55% to 88%.

After performing the first filter, we used filter II to reduce the number of complexes to around 2,000 to 4,000 models. The obtained results are shown in Table 1.

In Table 1, there are 11 complexes (1A0O, 1AHW, 1BRS, 1DFJ, 1FQ1, 1IGC, 1UDI, 1UGH, 1WQ1, 2MTA, 4HTC) for which we couldn't find sufficient homolog sequences from non-redundant databases to calculate the conserved residue position information. Therefore, only filter II is applied for these complexes.

When we applied the filters to the model sets, some near-native structures are also filtered out (false negatives), besides non-native structures. Here we define the improvement factor (I_fact) as I_fact = $(hits/models)_f$ / $(hits/models)_i$, where hits/models is the ratio of the number of structures with RMSD<4.5 Å from the native structure over the number of complex models, before ($(hits/models)_i$) and after ($(hits/models)_f$) applying the filters.

The results are shown in Table 1 and Figure 2. It is observed that there are 48 out of 60 complexes with I_fact greater than 1.0. Most of them (44) are greater than 2.0, which means the improvement is over 100%. For a few complexes applying the filter resulted in more than 400% improvement.

There are 5 out of 60 complexes (1AVW, 1BQL, 1EFU, 1FIN, 1GOT) with I_fact=1.0 for these 5 complexes (see section 2.1), FTDock did not generate any near-native structure (with RMSD less than 4.5Å), i.e. no hits are found. When we examined these structures more carefully, we found that except for 1FIN in which the LRMSD structure was filtered out, the LRMSD structures are still in the filtered subset of these proteins. Moreover, the filters have reduced the number of model structures for these five complexes by a factor

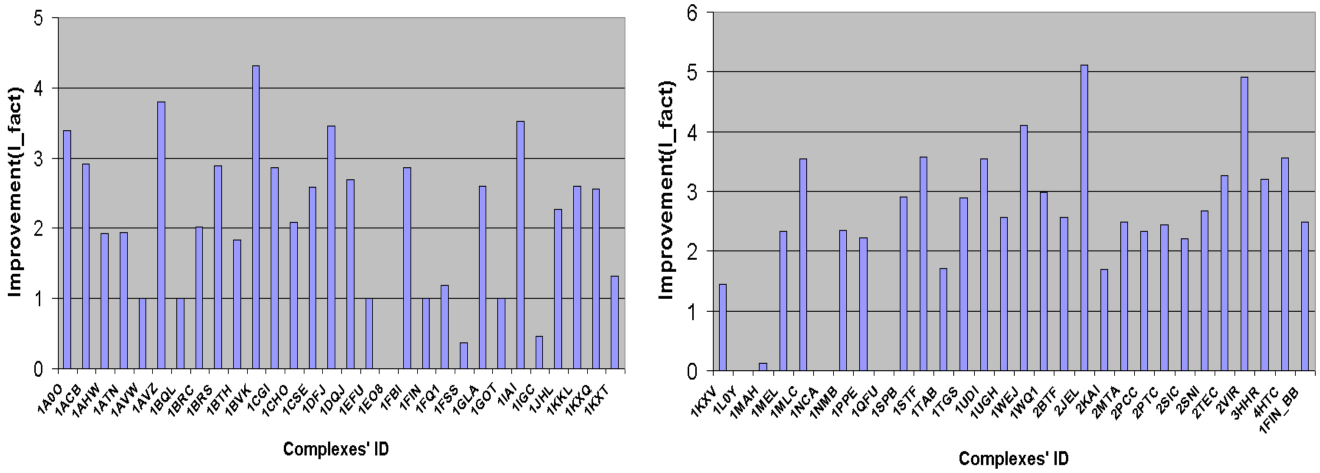

Figure 2. The improvement after filtering. The results are: (1) 48 out of 60 complexes have I_fact>1; (2) there are 5 complexes (1AVW, 1BQL, 1EFU, 1FIN, 1GOT) with I_fact=1 because FTDock did not generate hits to begin with; (3) There are 7 complexes for which our filters worsen the results (1FSS, 1IGC, 1MAH, 1EO8, 1L0Y, 1NCA, 1QFU) with 1>I_fact>=0 after filtering

of 2.5 to 4. This shows that the filters help with even these 5 complexes.

Our filters failed for 7 complexes: there are 3 complexes (1FSS, 1IGC, 1MAH) for which I_fact is less than 1.0 (Figure 2 and Table 1). For these structures proportionately more near-native model structures are filtered out than unrelated ones. In Figure 2, it can also be observed that 4 complexes (1EO8, 1L0Y, 1NCA, 1QFU) have I_fact=0. This means that we filtered out all of the near-native structures (2, 1, 7, and 5 hits for the four complexes respectively). When we examined the number of conserved residue positions at the interface for these 4 complexes we found that there is a high number of conserved residue positions for antibody-antigen systems 1EO8 and 1QFU, and a low number of conserved residues for non-antibody 1L0Y and 1NCA, contrary to most of the complexes investigated.

The global rank (see next section) for these 4 failed complexes (1EO8, 1L0Y, 1NCA, 1QFU) and two of the complexes (1FSS, 1MAH) without improvements are also given in Table 1 without using filter I. It is observed that except for 1L0Y, the I_fact values of the rest 5 complexes are greater than 1.0 and the lower RMSD models are still in the subset. 1L0Y only has one hit (see Table 3) and is filtered out by filter II, but other lower RMSD models are still in the subset. For 1IGC, since there are not enough homologous sequences from the database, we couldn't get the conserved residue position information. The result listed in Table 1 is obtained by just using filter II. Its improvement (I_fact) is still less than 1.0 since there are lower RMSD models filtered out.

3.3 The efficiency of global-ranking

The free energy of binding would in principle suffice to determine the native structure from a large set of complexes. Unfortunately, the free energy we calculated does not rank near native structures at the top of the list. This could be the result of inaccuracies in the potential force fields used for calculating enthalpic terms or in the empirical, entropic terms. Conformational changes upon binding whether local or global can also result in significant changes in the free energy of binding. As a result we have to resort to empirical descriptors, and since none can individually predict near-native structures with great accuracy, we decided to combine multiple descriptors in a global ranking scheme.

Empirical rankings based on more than one descriptor have been attempted before: in ZDOCK (Chen et al. 2003a) shape-complementarity, electrostatics and desolvation energies were combined to get a final target function, and AutoDock (Morris et al. 1998) involved more energy terms into the score function. A major bottleneck for composite, global scoring functions is that the weights for different quantities are hard to determine. As stated in section 2.5, we derived a global ranking function by re-normalizing the rank of each employed descriptor (eq.(6)). From our calculations, we obtained 6 quantities: shape complementarity, pair-potential, total internal energy from CHARMM minimization, binding free energy, desolvation energy, and conserved residue indices. From the correlation coeficients calculations we found that the pair potential descriptor has a significant correlation coeficient value (>0.10) for 22 complexes, that the desolvation energy has significant positive correlation in 13 complexes, that the conserved residue descriptor has significant correlation in 10 complexes, that shape complementarity values correlate well with RMSD in 3 complexes, and that binding free energy has significant correlation coefficient value in 3 complexes. In some complexes there are more than one descriptors with significant correlation coefficient values. Using the relative values of correlation coefficients we have derived weights 1, 1, 2, 4, and 5 for shape-complementarity, binding free energy, conservation index, desolvation energy and pair-potential energy respectively.

Using equation (6) and these weights we obtained a new global rank for each model complex. The rank of the LRMSD structure for each complex is also listed in Table 1 (G_rank). From it, we can see that in most of the model complexes the near native complexes have lower ranks. But in some cases, with higher RMSD have lower ranks (false positive).

In Table 1, we also give the number of hits (E_hits) within the first 100 ranks. For 22 complexes, application of the global ranking resulted in no hits in the top 100 ranked structures. We should note that for five of these there were no hits to begin with, because FTDock did not generate any.

For the rest 38 complexes, application of the global ranking improves substantially the predictive ability. Specifically, we calculate the improvement over random (IOR) for these 38 complexes (IOR=(E_hits/100)/(Hits/NRC)), where NRC is the number of complexes after filtering, and we find substantial IOR values. The average calculated IOR for these 38 complexes is 11.18. Even when the 17 complexes with IOR=0 are included in the average calculation, the average IOR for the 55 complexes that FTDock generated hits is 7.72.

4. Concluding remarks

In this work we have demonstrated the usefulness of conserved residue position information in identifying possible near-native complex model structures from docking solutions. We have used this information to develop two filters, reducing the number of docked model structures by 55% to 88% depending on the complex, while keeping near-native complexes in the remaining subset. We applied our method to a benchmark set of 59 complexes. There are 11 complexes for which we didn't find enough homolog sequence information. Thus, we could not apply our filter at present. Only for 4 out of the rest complexes our filter failed to retain the near-native structures, and for another 3 out of 60 complexes (the 59 benchmark and the FIN bound-bound calculation) our filter did poorly compaired to FTDock results. After filtering, we minimized the side-chain structure of the remaining model structures, and we calculated the binding free energy and desolvation energy. We developed a ranking scheme by renormalizing and weighting a combination of the ranks based on conservation position information, shape complementarity, desolvation energy, pair potential and binding free energy. Excluding the five complexes for which FTDock did not generate any hits (with RMSD<4.5Å), the average improvement over random for the top 100 ranked structures is 7.72. For 17 complexes IOR=0, but for the majority (38 complexes) we observed significant improvements in predictive ability, in terms of predicting near-native structures in the highest-ranked 100 structures.

Table 1. The number of complexes (NRC) after applying filters. G_rank: Sorting global score calculated from equation (6). I_fact: improvement factor by our filter. E_hits: Number of hits within the first 100 ranks. IOR: Improvement over random. The numbers in italic are obtained by performing only filter II.

Complex	NRC	LRMS(A)	Hits	I_fact	IOR	Complex	NRC	LRMSD(A)	Hits	I_fact	IOR
		/(G_rank)	/(E_hits)					/(G_rank)	/(E_hits)		
1A0O	2721	2.62(31)	12(4)	3.39	9.07	**1KXV**	3901	1.69(139)	9(0)	1.44	0
						1L0Y	3004	4.87	0	0	
1ACB	2485	0.31(89)	21(15)	2.91	17.75		*2799*	*4.87(1587)*	*0(0)*	*0*	*0*
						1MAH	3258	3.31	2	0.13	
1AHW	*2875*	*3.37(316)*	*5(0)*	*1.93*	*0*		*2754*	*1.11(653)*	*20(2)*	*1.58*	*2.75*
1ATN	4211	0.40(7)	9(7)	1.94	32.75	**1MEL**	2579	1.26(7)	6(3)	2.33	12.90
1AVW	3998	4.69(2102)	0(0)	1	0	**1MLC**	2544	1.39(58)	9(3)	3.54	8.48
						1NCA	1637	4.95	0	0	
1AVZ	2210	2.64(1686)	21(0)	3.80	0		*2778*	*0.41(600)*	*4(0)*	*2.06*	*0*
1BQL	*2615*	*5.30(570)*	*0(0)*	*1*	*0*	**1NMB**	2136	0.77(148)	3(0)	2.34	0
1BRC	2825	1.29(12)	12(6)	2.02	14.1	**1PPE**	2695	0.56(5)	394(74)	2.22	5.00
						1QFU	1439	5.33	0	0	
1BRS	*2785*	*1.66(13)*	*29(9)*	*2.89*	*8.64*		*2781*	*0.70(190)*	*5(3)*	*3.60*	*16.6*
1BTH	3110	3.70(693)	8(0)	1.84	0	**1SPB**	2837	0.95(1)	33(9)	2.91	7.73
1BVK	2322	3.03(512)	13(1)	4.31	1.79	**1STF**	2803	0.63(2)	15(10)	3.57	18.67
1CGI	2753	1.58(84)	70(25)	2.86	9.83	**1TAB**	3591	0.72(1318)	49(0)	1.71	0
1CHO	3504	2.23(21)	44(19)	2.09	15.1	**1TGS**	3046	1.64(54)	22(11)	2.89	15.23
1CSE	2981	0.92(304)	43(5)	2.58	3.47	**1UDI**	*2824*	*2.34(18)*	*24(3)*	*3.54*	*3.53*
1DFJ	*2894*	*3.24(997)*	*1(0)*	*3.46*	*0*	**1UGH**	*2785*	*3.80(771)*	*5(1)*	*2.56*	*5.57*
1DQJ	3428	2.95(1306)	12(0)	2.69	0	**1WEJ**	2433	2.74(451)	10(0)	4.11	0
1EFU	4205	5.71(898)	0(0)	1	0	**1WQ1**	*2833*	*2.39(510)*	*11(0)*	*2.99*	*0*
1EO8	1395	4.92	0	0							
	2812	*3.01(134)*	*2(1)*	*3.56*	*14.06*	**2BTF**	3909	1.61(59)	5(4)	2.56	31.27
1FBI	2880	2.68(1381)	14(0)	2.86	0	**2JEL**	1955	2.86(40)	8(3)	5.12	7.33
1FIN	4072	5.94(389)	0(0)	1	0	**2KAI**	2801	1.51(154)	36(4)	1.69	3.11
1FQ1	*2810*	*3.05(379)*	*2(0)*	*1.19*	*0*	**2MTA**	*2676*	*2.87(4)*	*10(1)*	*2.49*	*2.68*
1FSS	2694	3.80	7	0.37							
	2697	*1.81(1086)*	*42(2)*	*2.26*	*1.28*	**2PCC**	3410	2.28(500)	31(2)	2.33	2.20
1GLA	3846	2.75(34)	5(3)	2.60	23.18	**2PTC**	2914	1.42(71)	42(10)	2.44	6.94
1GOT	3245	5.80(222)	0(0)	1	0	**2SIC**	3690	1.86(1)	13(11)	2.20	31.22
1IAI	2274	3.27(29)	4(1)	3.52	5.68	**2SNI**	2928	2.52(268)	18(5)	2.67	8.13
1IGC	*2628*	*1.84(2552)*	*11(0)*	*0.46*	*0*	**2TEC**	2943	0.45(139)	52(11)	3.27	6.22
1JHL	4404	0.74(1414)	19(0)	2.27	0	**2VIR**	2035	0.80(362)	5(2)	4.91	8.14
1KKL	3076	3.32(2569)	4(0)	2.60	0	**3HHR**	3125	4.50(889)	1(0)	3.20	0
1KXQ	3476	0.46(1)	8(3)	2.56	13.04	**4HTC**	*2813*	*1.46(1)*	*11(6)*	*3.56*	*15.34*
1KXT	3794	0.45(265)	8(0)	1.32	0	**1FIN_BB**	4035	0.41(2)	10(9)	2.48	36.32

Acknowledgments

This work was supported in part by the Minnesota Supercomputer Institute (MSI) and the University of Minnesota Biotechnology Institute. This work is also supported by the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, Army Research Laboratory, contract number DAAD10-01-2-0014. The content does not necessarily reflect the position or the policy of the government and no official endorsement *should be inferred. AHPCRC and the MSI provided access to computing facilities.*

References

Brooks, B.R., Bruccoleri, R.E., Olfson, B.D., States, D.J., Swaminathan, S., and Karplus, K. 1983. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J Comp Chem* **4:** 187-217.

Camacho, C.J., and Vajda, S. 2002. Protein-protein association kinetics and protein docking. *Curr Opin Struct Biol* **12:** 36-40.

Chen, R., Li, L., and Weng, Z. 2003a. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52:** 80-87.

Chen, R., Mintseris, J., Janin, J., and Weng, Z. 2003b. A protein-protein docking benchmark. *Proteins* **52:** 88-91.

Cherfils, J., and Janin, J. 1993. Protein docking algorithms: Simulating molecular recgnition. *Curr. Opin. Struct. Biol.* **3:** 265-269.

Dennis, S., Kortvelyesi, T., and Vajda, S. 2002. Computational mapping identifies the binding sites of organic solvents on proteins. *Proc Natl Acad Sci U S A* **99:** 4290-4295.

Devore, J., and Peck, R. 2001. *Statistics: The Exploration and Analysis of Data*, 4th ed. Duxbury, pp. 136.

Dominguez, C., Boelens, R., and Bonvin, A.M. 2003. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125:** 1731-1737.

Ewing, T.J., Makino, S., Skillman, A.G., and Kuntz, I.D. 2001. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **15:** 411-428.

Fahmy, A., and Wagner, G. 2002. TreeDock: a tool for protein docking based on minimizing van der Waals energies. *J Am Chem Soc* **124:** 1241-1250.

Gabb, H.A., Jackson, R.M., and Sternberg, M.J. 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* **272:** 106-120.

Gardiner, E.J., Willett, P., and Artymiuk, P.J. 2003. GAPDOCK: a Genetic Algorithm Approach to Protein Docking in CAPRI round 1. *Proteins* **52:** 10-14.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19:** 163-164.

Gonnet, G.H., Cohen, M.A., and Benner, S.A. 1992. Exhaustive Matching of the Entire Protein Sequence Database. *Science* **256:** 1443-1445.

Gottschalk, K.E., Neuvirth, H., and Schreiber, G. 2004. A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. *Protein Eng Des Sel* **17:** 183-189.

Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A., and Baker, D. 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* **331:** 281-299.

Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47:** 409-443.

Helmer-Citterich, M., and Tramontano, A. 1994. PUZZLE: a new method for automated protein docking based on surface shape complementarity. *J Mol Biol* **235:** 1021-1031.

Janin, J. 1995. Protein-protein recognition. *Prog Biophys Mol Biol* **64:** 145-166.

Jayaram, B., McConnell, K.J., Dixit, S.B., and Beveridge, D.L. 1999. Free energy analysis of protein-DNA binding: The EcoRI endonuclease-DNA complex. *J Comput Phys* **151:** 333-357.

Jayaram, B., Sprous, D., and Beveridge, D.L. 1998. Solvation free energy of biomacromolecules: Parameters for a modified generalized born model consistent with the AMBER force field. *J Phys Chem B* **102:** 9571-9576.

Jones, S., and Thornton, J.M. 1996. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93:** 13-20.

Jones, S., and Thornton, J.M. 1997. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* **272:** 133-143.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., and Vakser, I.A. 1992. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A* **89:** 2195-2199.

Larsen, T.A., Olson, A.J., and Goodsell, D.S. 1998. Morphology of protein-protein interfaces. *Structure* **6:** 421-427.

Lichtarge, O., and Sowa, M.E. 2002. Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struct Biol* **12:** 21-27.

Lo Conte, L., Chothia, C., and Janin, J. 1999. The atomic structure of protein-protein recognition sites. *J Mol Biol* **285:** 2177-2198.

Mandell, J.G., Roberts, V.A., Pique, M.E., Kotlovyi, V., Mitchell, J.C., Nelson, E., Tsigelny, I., and Ten Eyck, L.F. 2001. Protein docking using continuum electrostatics and geometric fit. *Protein Eng* **14:** 105-113.

Moont, G., Gabb, H.A., and Sternberg, M.J. 1999. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* **35:** 364-373.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., and Olson, A.J. 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **19:** 1639-1662.

Palma, P.N., Krippahl, L., Wampler, J.E., and Moura, J.J. 2000. BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins* **39:** 372-384.

Paul, N., and Rognan, D. 2002. ConsDock: A new program for the consensus analysis of protein-ligand interactions. *Proteins* **47:** 521-533.

Reddy, B.V.B., and Kaznessis, Y. 2004. A quantitive analysis of interfacial amino acid conservation in protein-protein hetro complex structures. *Submitted to Appl. Bioinformatics*.

Richmond, T.J., and Richards, F.M. 1978. Packing of alpha-helices: geometrical constraints and contact areas. *J Mol Biol* **119:** 537-555.

Ritchie, D.W., and Kemp, G.J. 2000. Protein docking using spherical polar Fourier correlations. *Proteins* **39:** 178-194.

Sali, A., and Blundell, T.L. 1990. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol* **212:** 403-428.

Shoichet, B.K., and Kuntz, I.D. 1996. Predicting the structure of protein complexes: a step in the right direction. *Chem Biol* **3:** 151-156.

Smith, G.R., and Sternberg, M.J. 2002. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol* **12:** 28-35.

Sternberg, M.J., Gabb, H.A., and Jackson, R.M. 1998. Predictive docking of protein-protein and protein-DNA complexes. *Curr Opin Struct Biol* **8:** 250-256.

Taylor, J.S., and Burnett, R.M. 2000. DARWIN: a program for docking flexible molecules. *Proteins* **41:** 173-191.

Tsai, C.J., Lin, S.L., Wolfson, H.J., and Nussinov, R. 1997. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci* **6:** 53-64.

Valdar, W.S., and Thornton, J.M. 2001. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42:** 108-124.

Wang, Y.H., Zhang, H., and Scott, R.A. 1995. A New Computational Model for Protein-Folding Based on Atomic Solvation. *Protein Science* **4:** 1402-1411.

Yang, J.M., and Chen, C.C. 2004. GEMDOCK: a generic evolutionary method for molecular docking. *Proteins* **55:** 288-304.